Predicting User Satisfaction from Medicine Data: An In-Depth Analysis

Anumol Issac

Objectives

- **Analyze factors influencing user satisfaction:**
  - Explore the dataset to identify key features that impact user satisfaction.
- **Predict user satisfaction Rating:**
  - Apply machine learning models to predict user satisfaction rating based on the identified features and data insights.
- **Deploy the model:**
  - Select the best performing model and deploy it using Streamlit for practical usability and user interaction.

Data Summary

- **Dataset Overview**:

  Total records: 11825

  Total variables: 9

  Missing values: No missing values

- **Data source**: Kaggle

```
*****Dataset Overview*****

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11825 entries, 0 to 11824
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Medicine Name         11825 non-null  object
 1   Composition           11825 non-null  object
 2   Uses                  11825 non-null  object
 3   Side_effects          11825 non-null  object
 4   Image URL             11825 non-null  object
 5   Manufacturer          11825 non-null  object
 6   Excellent Review %    11825 non-null  int64
 7   Average Review %      11825 non-null  int64
 8   Poor Review %         11825 non-null  int64
dtypes: int64(3), object(6)
memory usage: 831.6+ KB
None
```

Analytical Approach

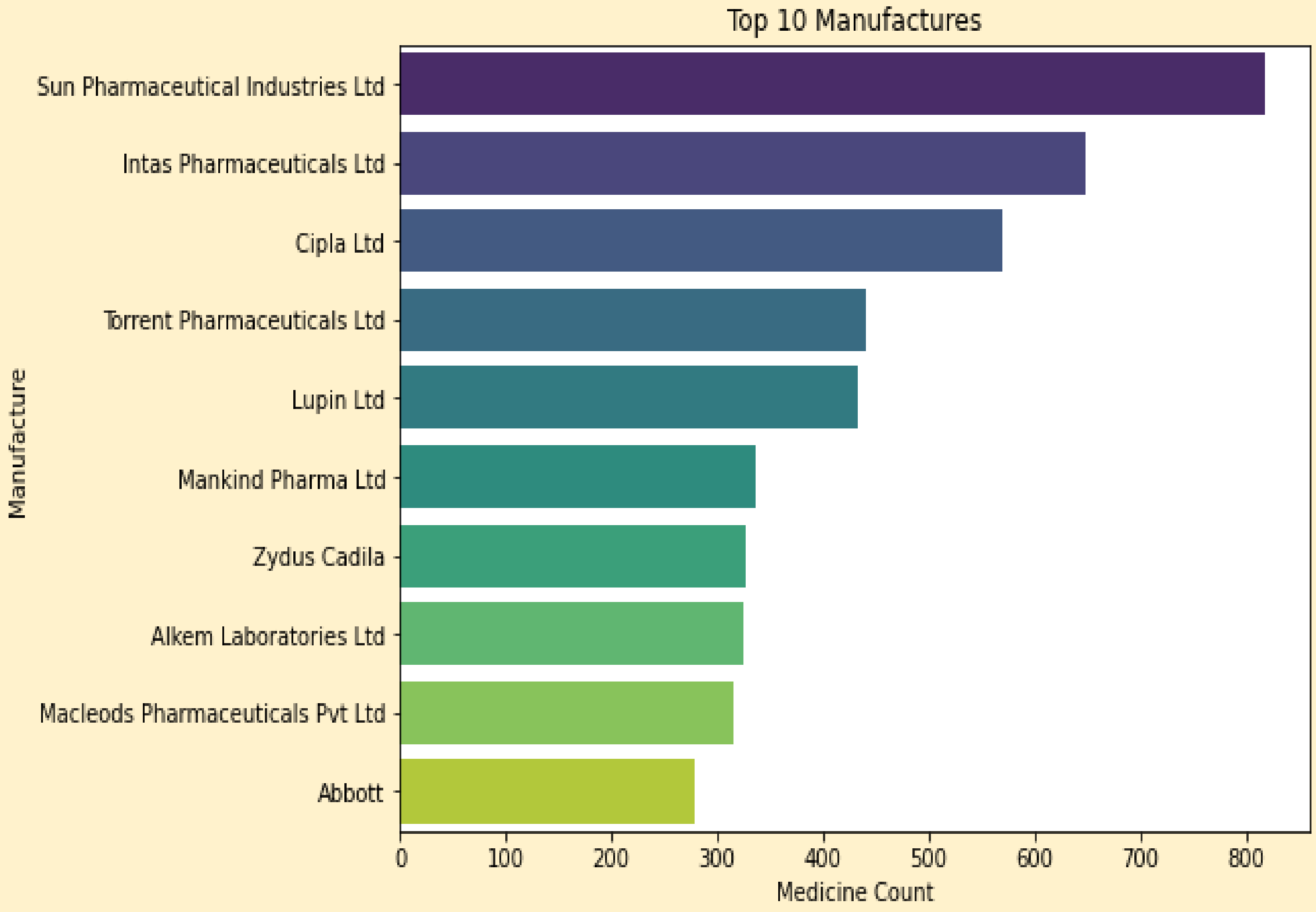
- **Data Exploration and Preprocessing:**
  - Clean and preprocess the dataset, handling missing values and duplicate entries.
  - Create new feature 'Rating' based on review percentages.

Rating = ((5* ['Excellent Review %'] + 3* ['Average Review %'] + 1* ['Poor Review %'])/100)

- **Understanding Features:**
  - Analyze the data through visualization and understand the relationship between different features.
- **Encoding techniques:**
  - Apply encoding techniques to convert categorical features into numerical form.
- **Feature Engineering and Target Set Creation:**
  - Create a feature and target set for prediction by selecting relevant input variables and defining the target variable.
- **Train-Test Split:**
  - Split the dataset into training and testing sets.
- **Apply Machine Learning Techniques:**
  - Apply machine learning techniques to predict user satisfaction and compare performance of different models based on metrics like MSE and R2.
- **Deploy the best model:**
  - Select the best-performing model based on evaluation metrics and deploy it for future predictions.

Top 10 Manufactures By Product Count

- **Sun Pharmaceutical Industries Ltd** has the highest product count of 819.
- **Intas Pharmaceuticals Ltd** follows with 648 products.
- **Cipla Ltd** ranks third with 569 products

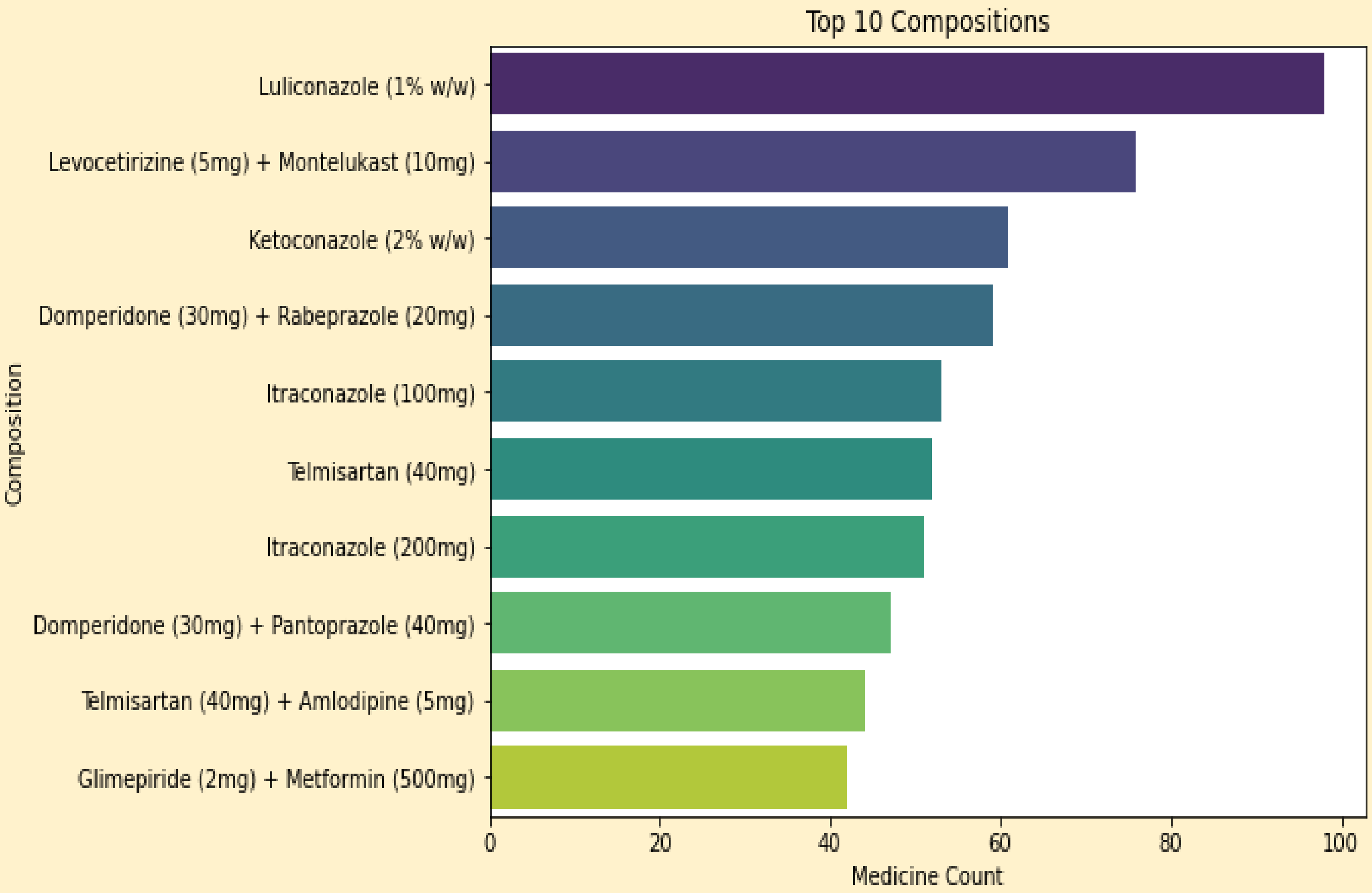


Top 10 Compositions

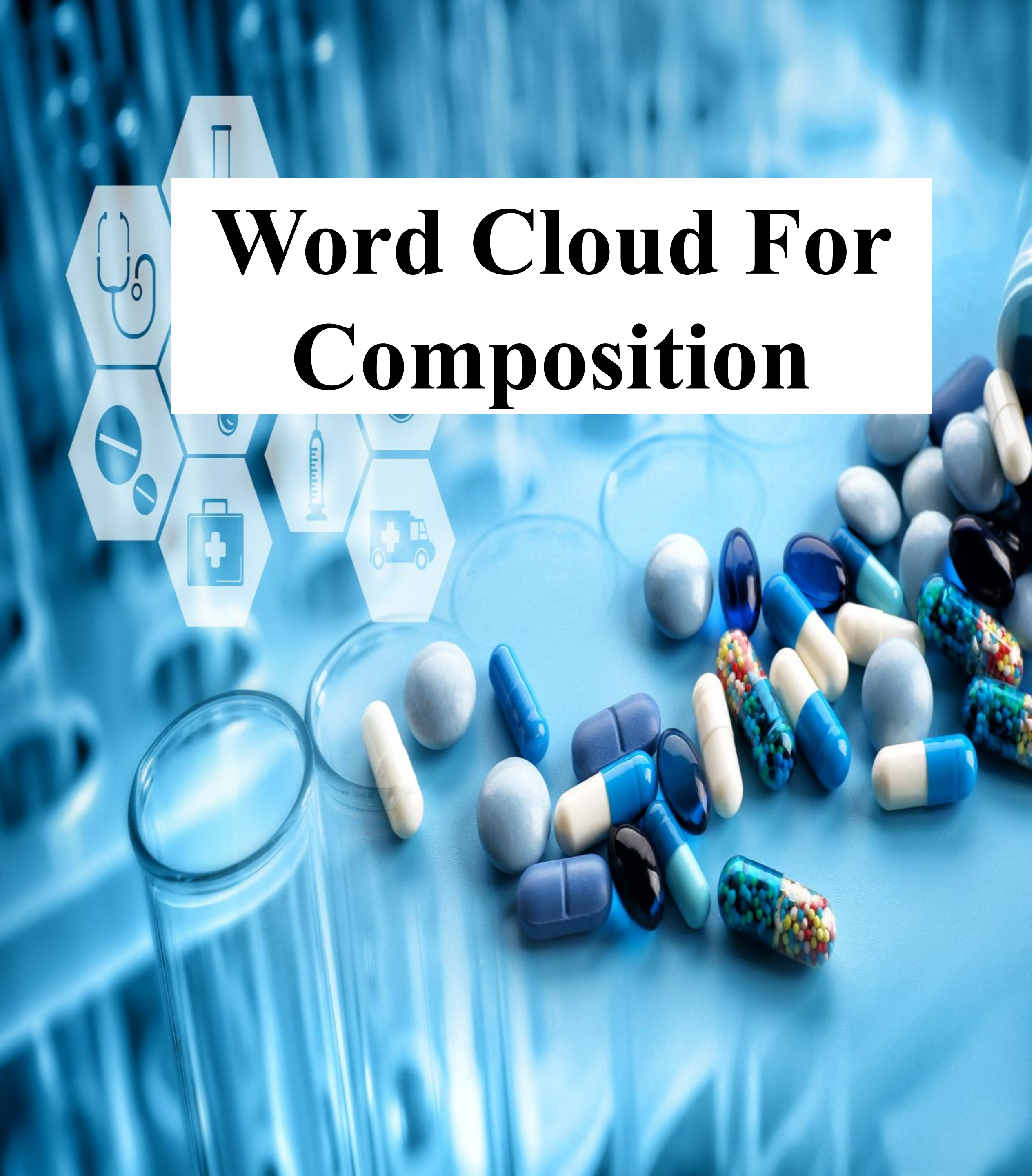
- **Luliconazole (1% w/w)** appears in the highest number of medicines (98 products).

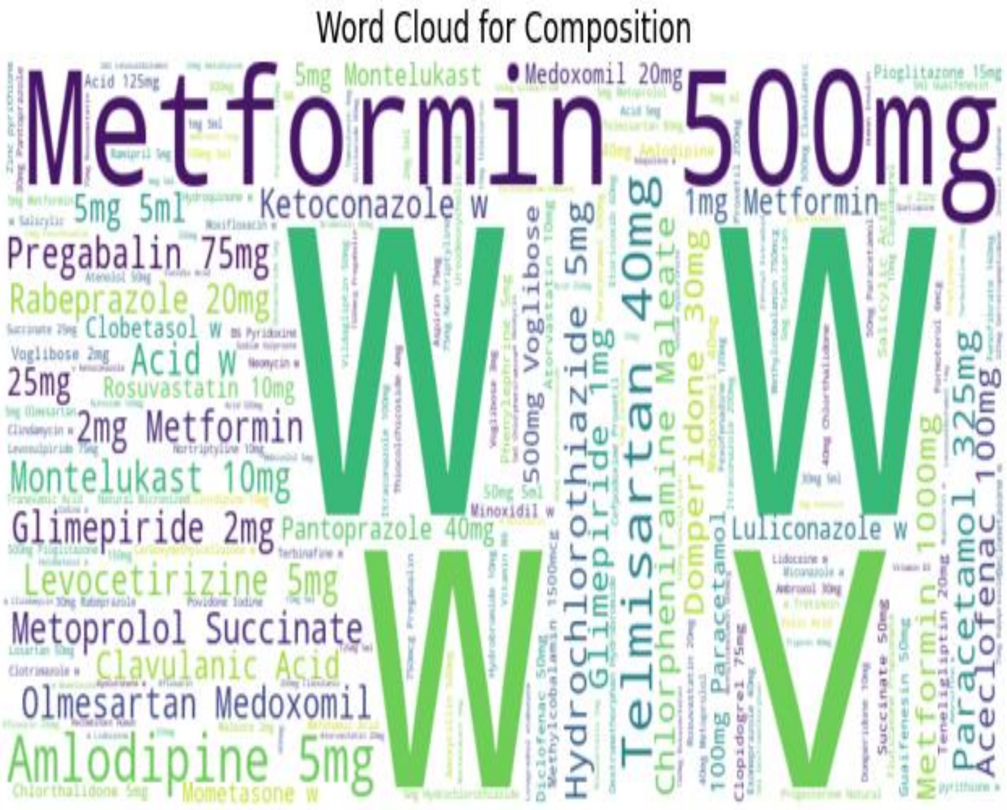
- **Levocetirizine (5mg) + Montelukast (10mg)** is the second most common composition, found in 76 medicines.

- **Ketoconazole (2% w/w)** and **Domperidone (30mg) + Rabeprazole (20mg)** are present in 61 and 59 medicines, respectively.

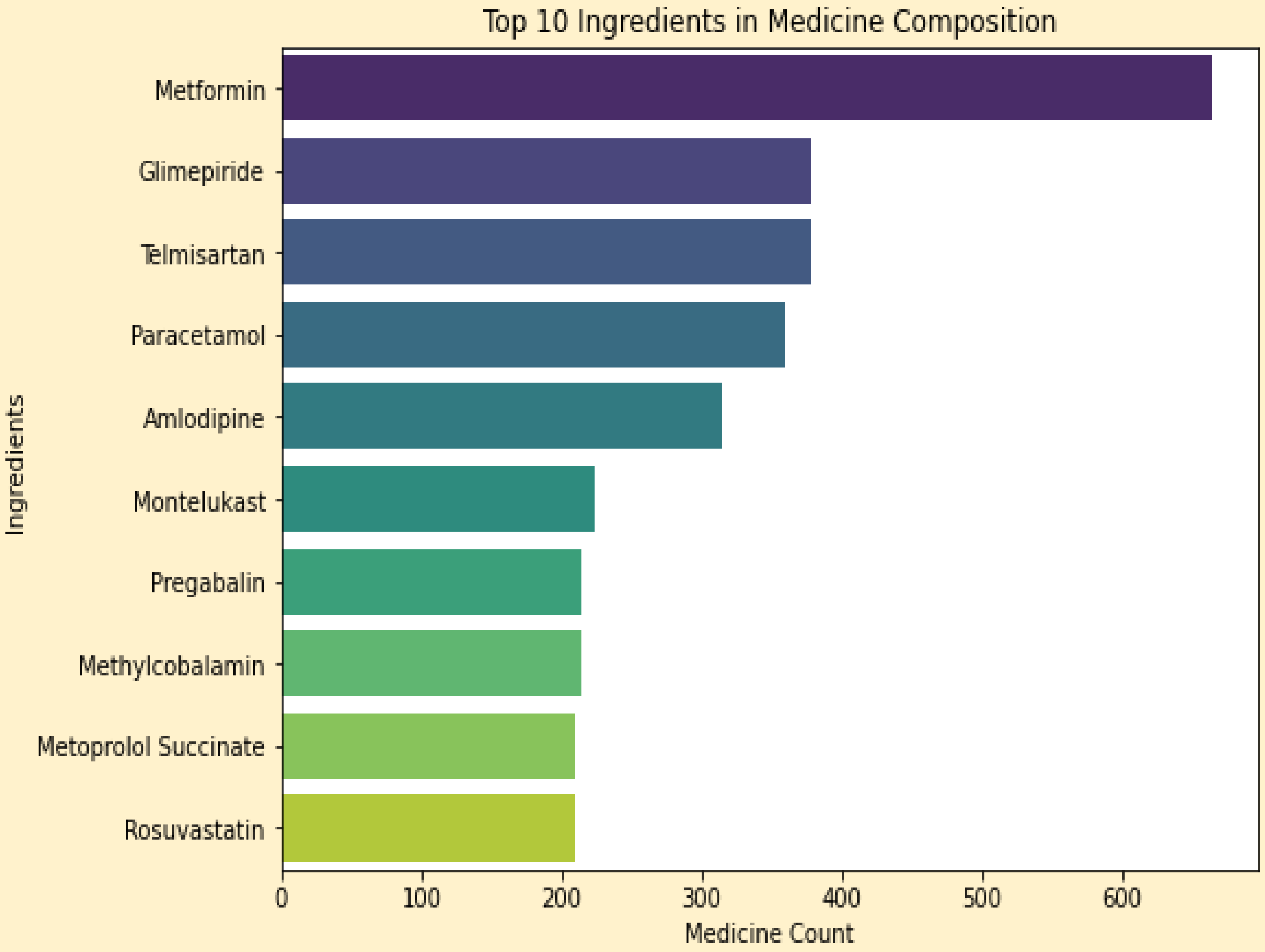


Word Cloud For Composition



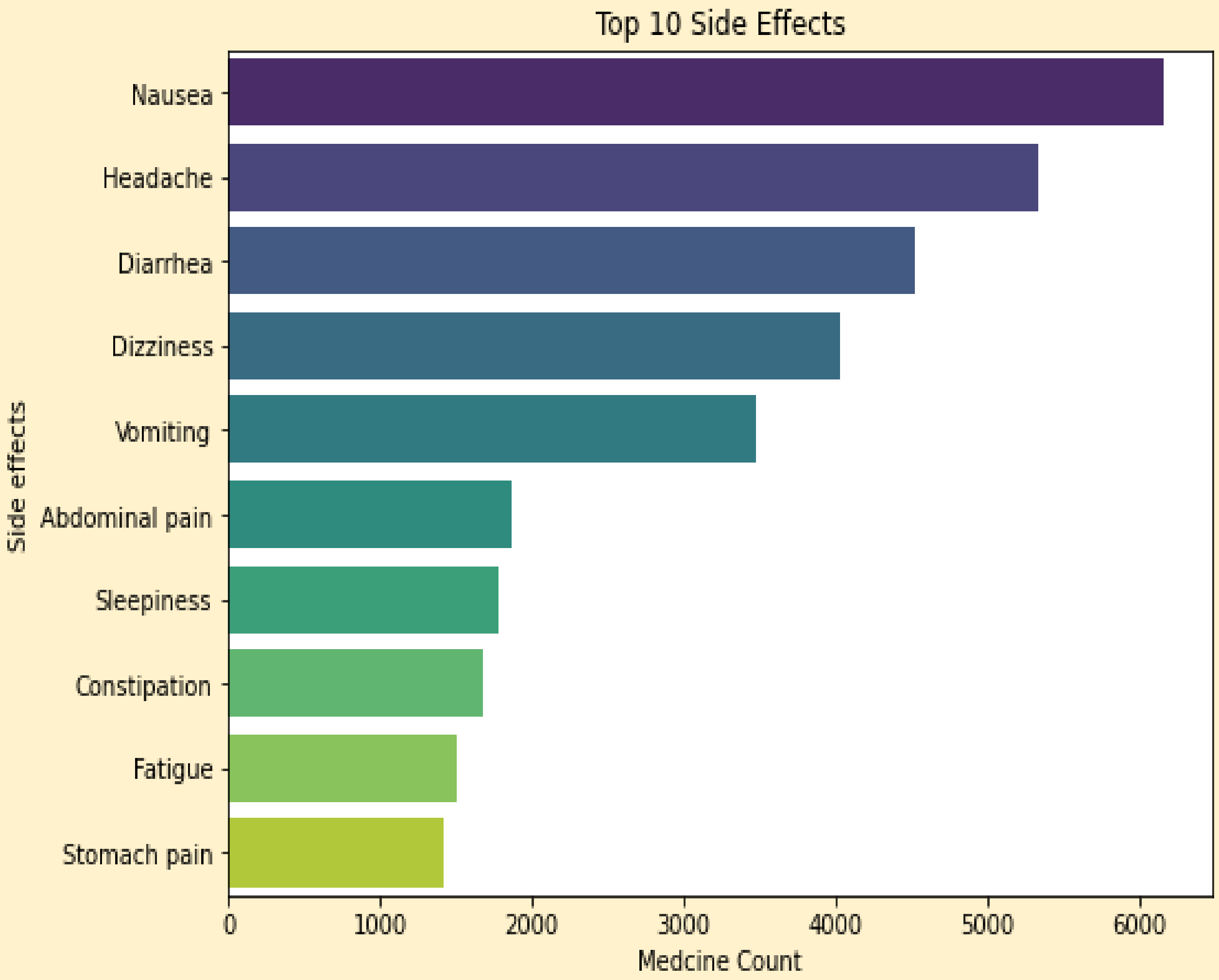


Most Commonly Used(top 10) Ingredients In Medicine Composition



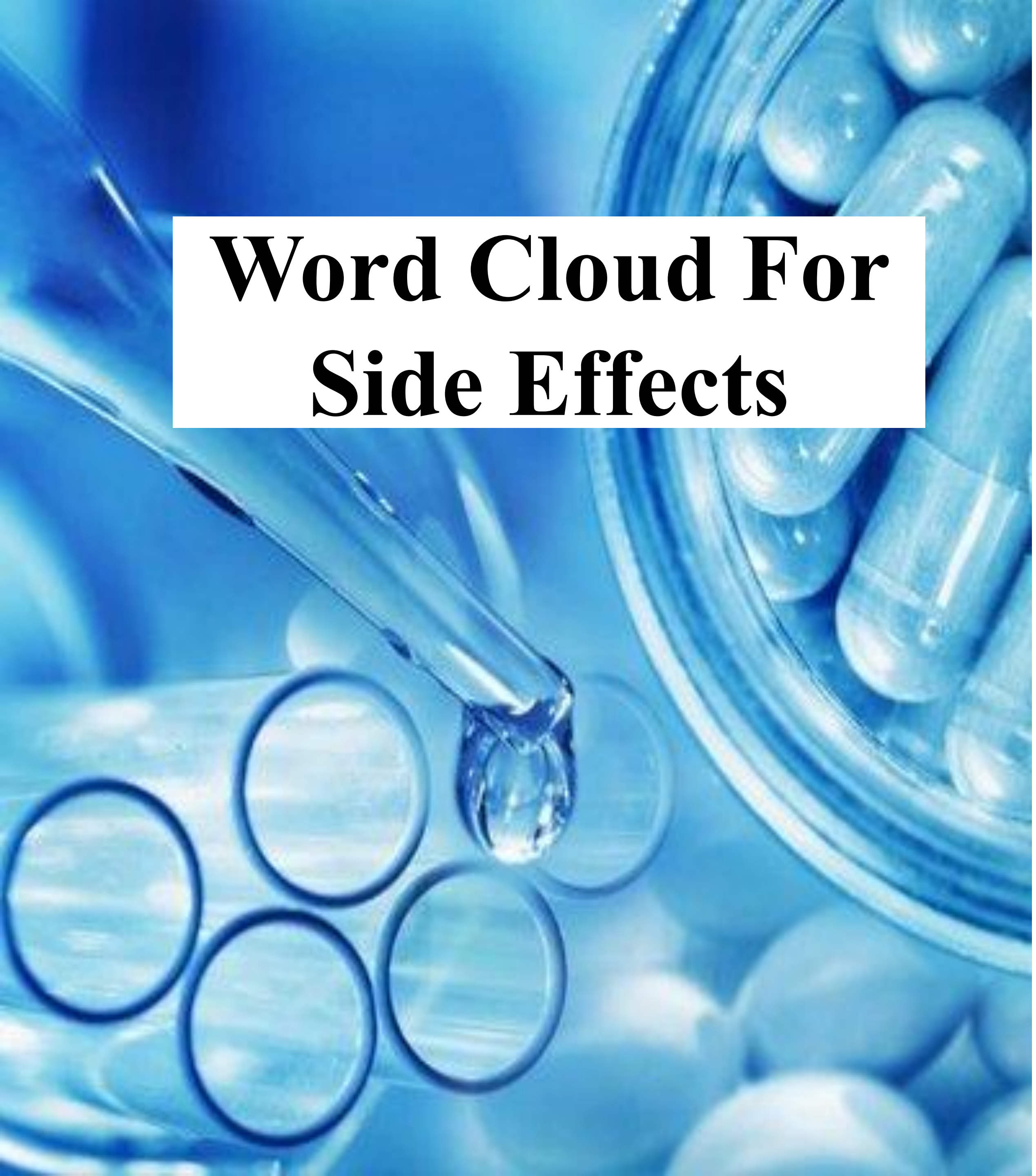
- ❑ **Metformin** is the most common ingredient, present in 664 medicines, commonly used for diabetes management.
- ❑ **Glimepiride** and **Telmisartan** are the next most frequent ingredients, appearing in 379 medicines each.
- ❑ **Paracetamol** ranks fourth, found in 359 medicines

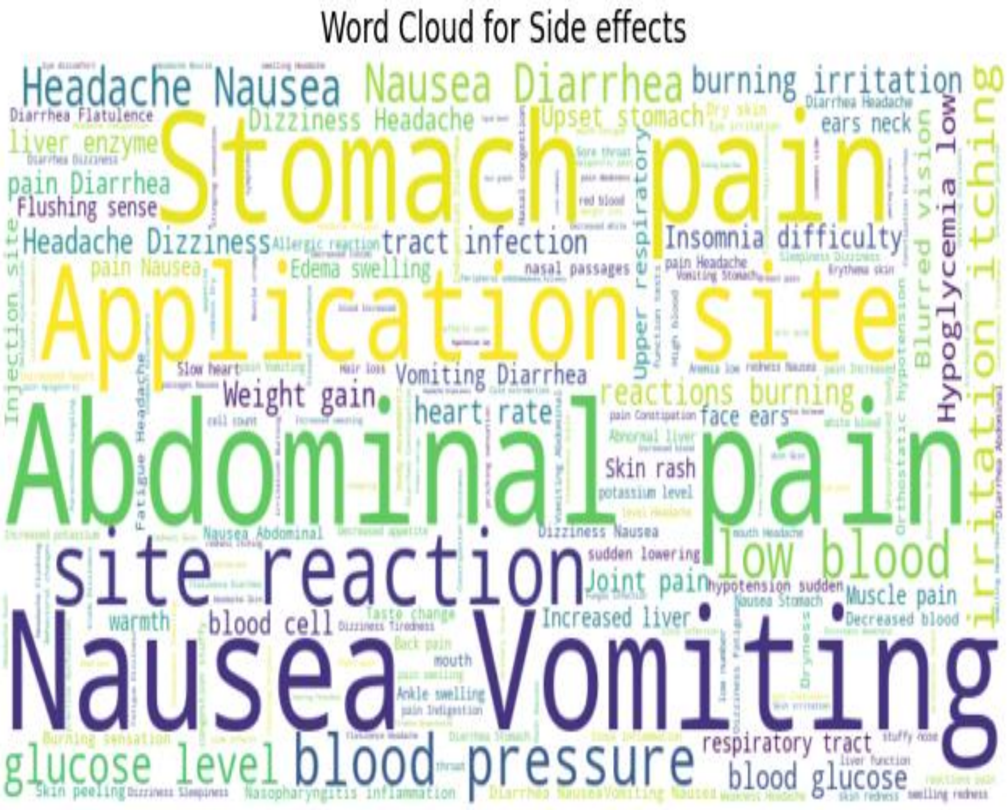
Top 10 (Most Common) Side Effects



- **Nausea** is the most frequently reported side effect, found in **6,170** medicines.
- **Headache** occurs in **5,336** medicines, making it the second most common side effect.
- **Diarrhea** is reported in **4,520** medicines.

Word Cloud For Side Effects



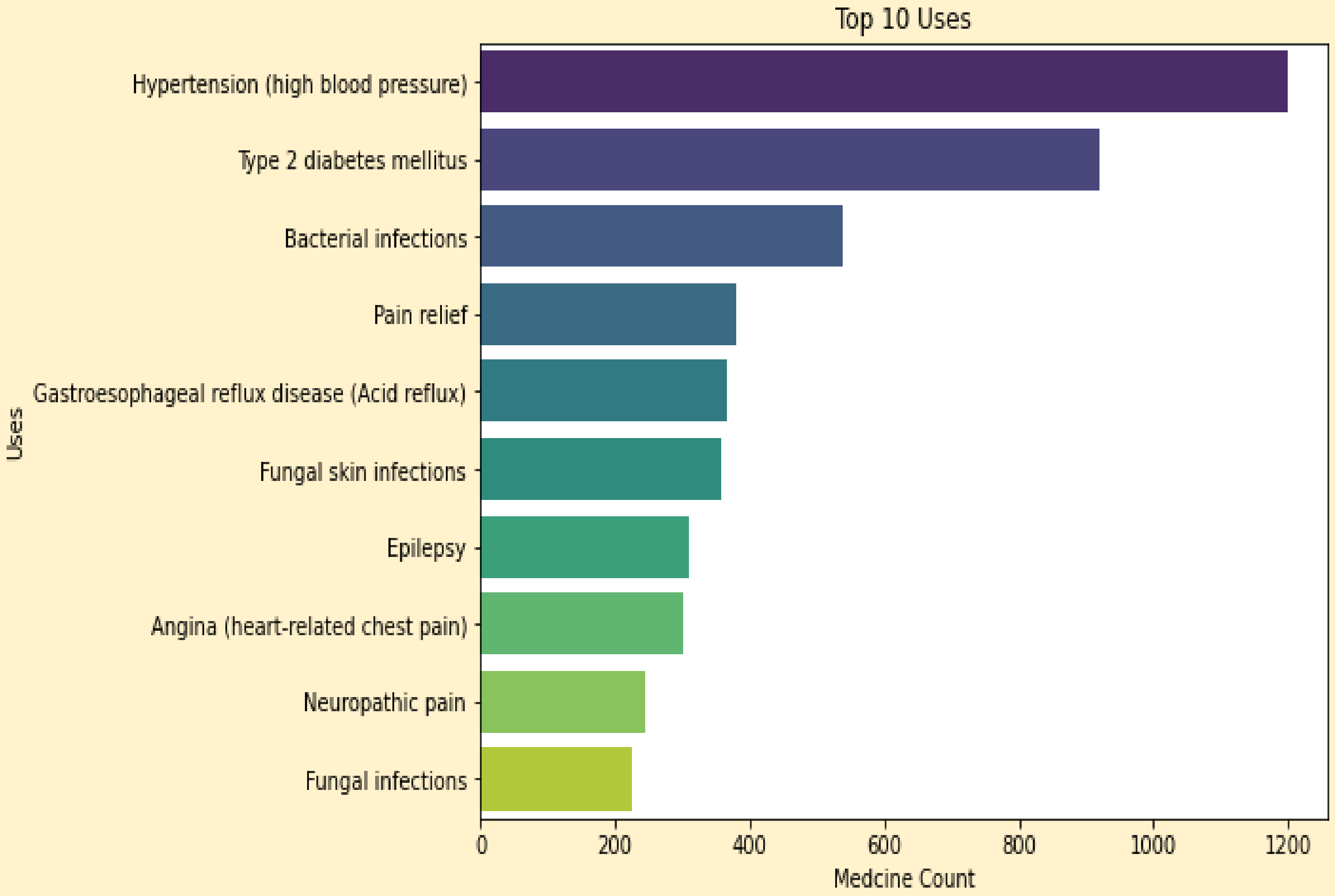


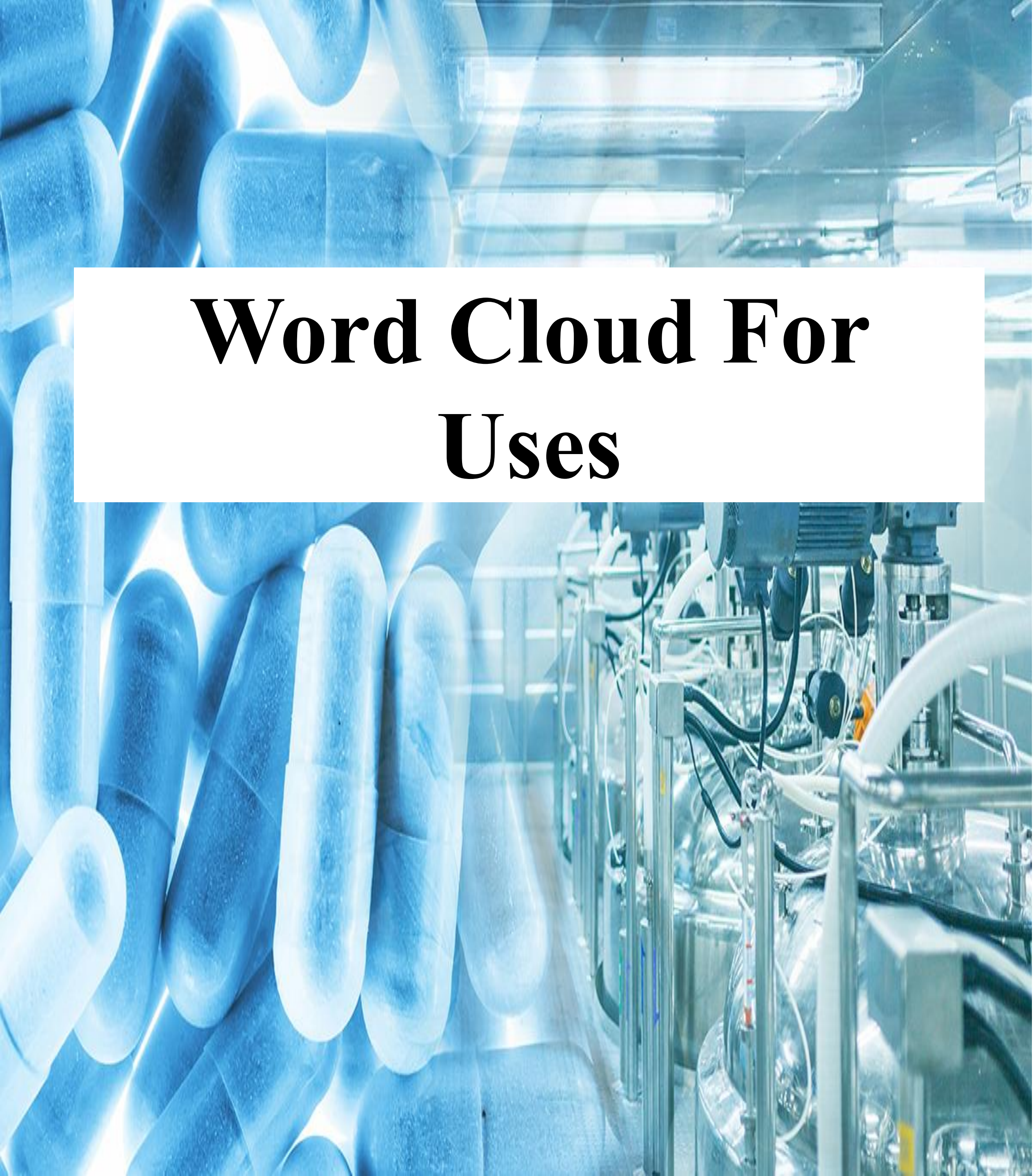
Top 10 Uses

- **Treatment of hypertension (high blood pressure)** is the most common use, with 1200 medicines used for this condition.

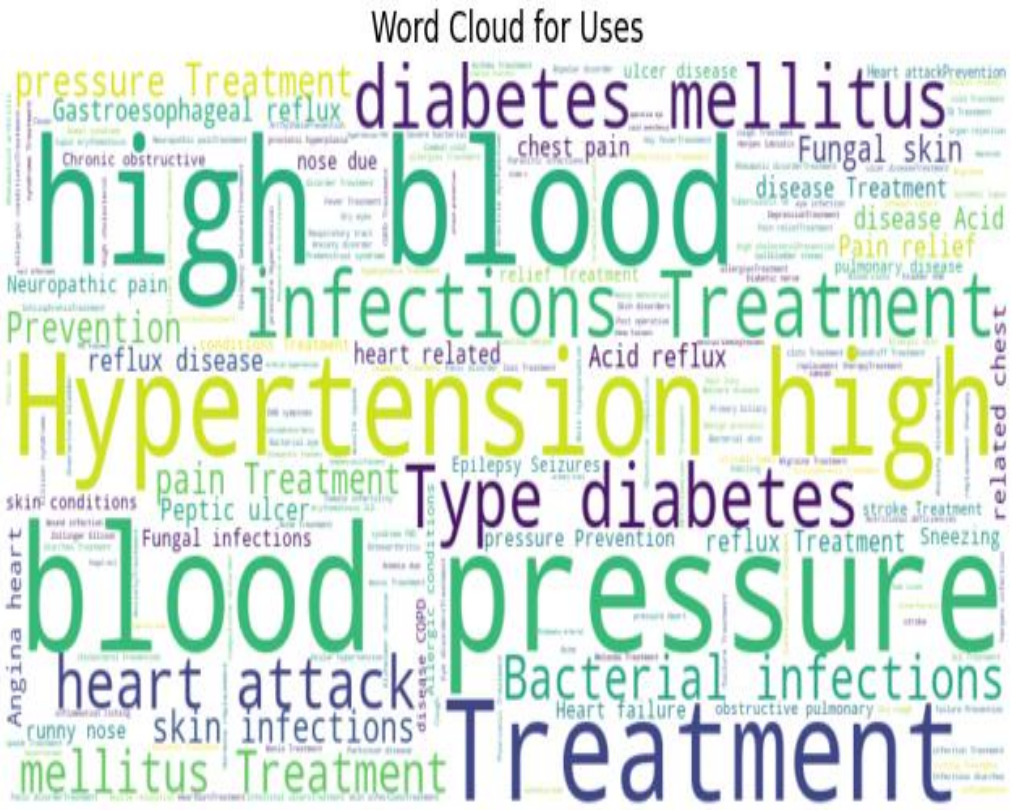
- **Treatment of type 2 diabetes mellitus** is treated by **920** medicines.

- **Treatment of bacterial infections** are third most common use, with **540** medicines used for this condition.

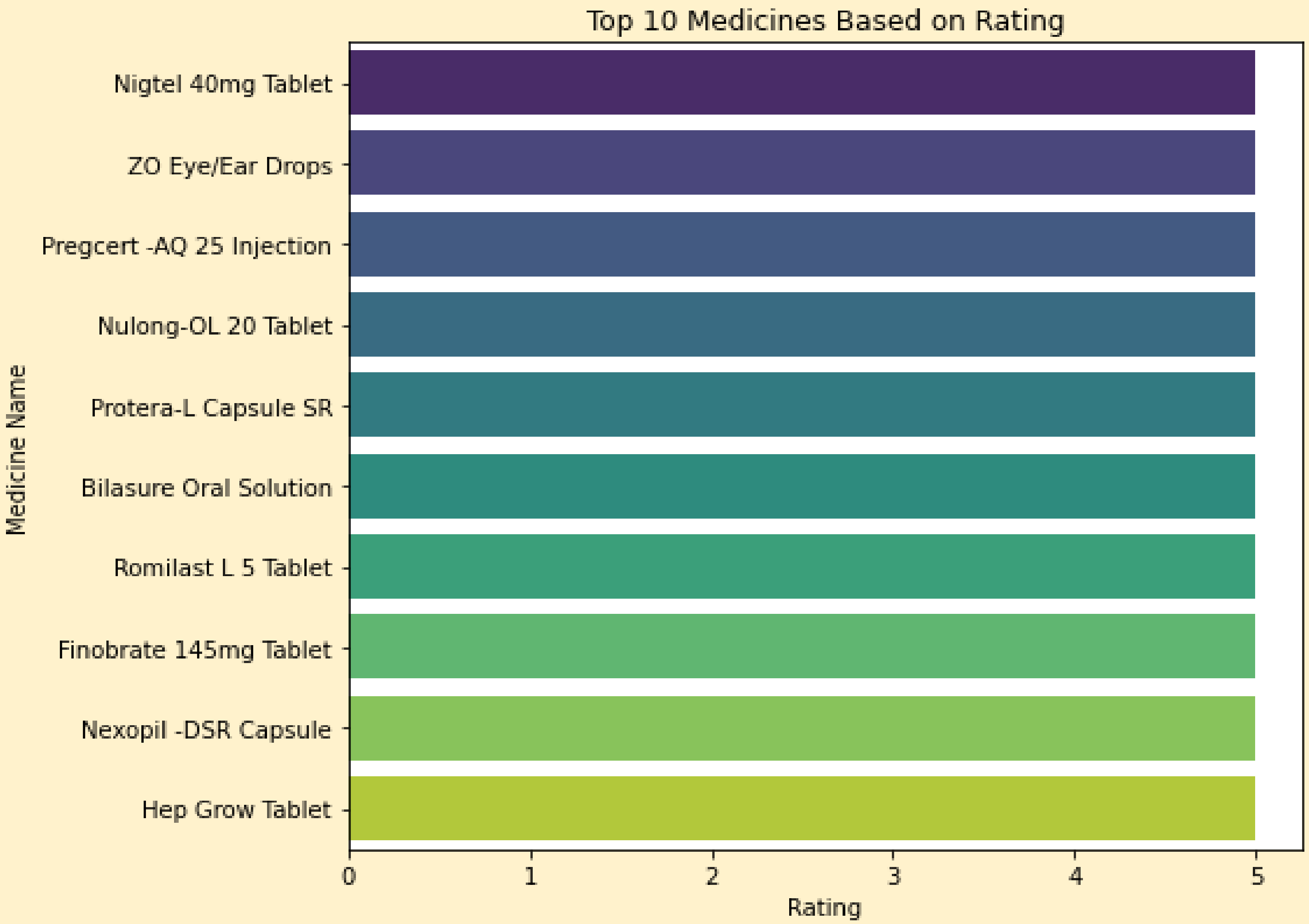




Word Cloud For Uses

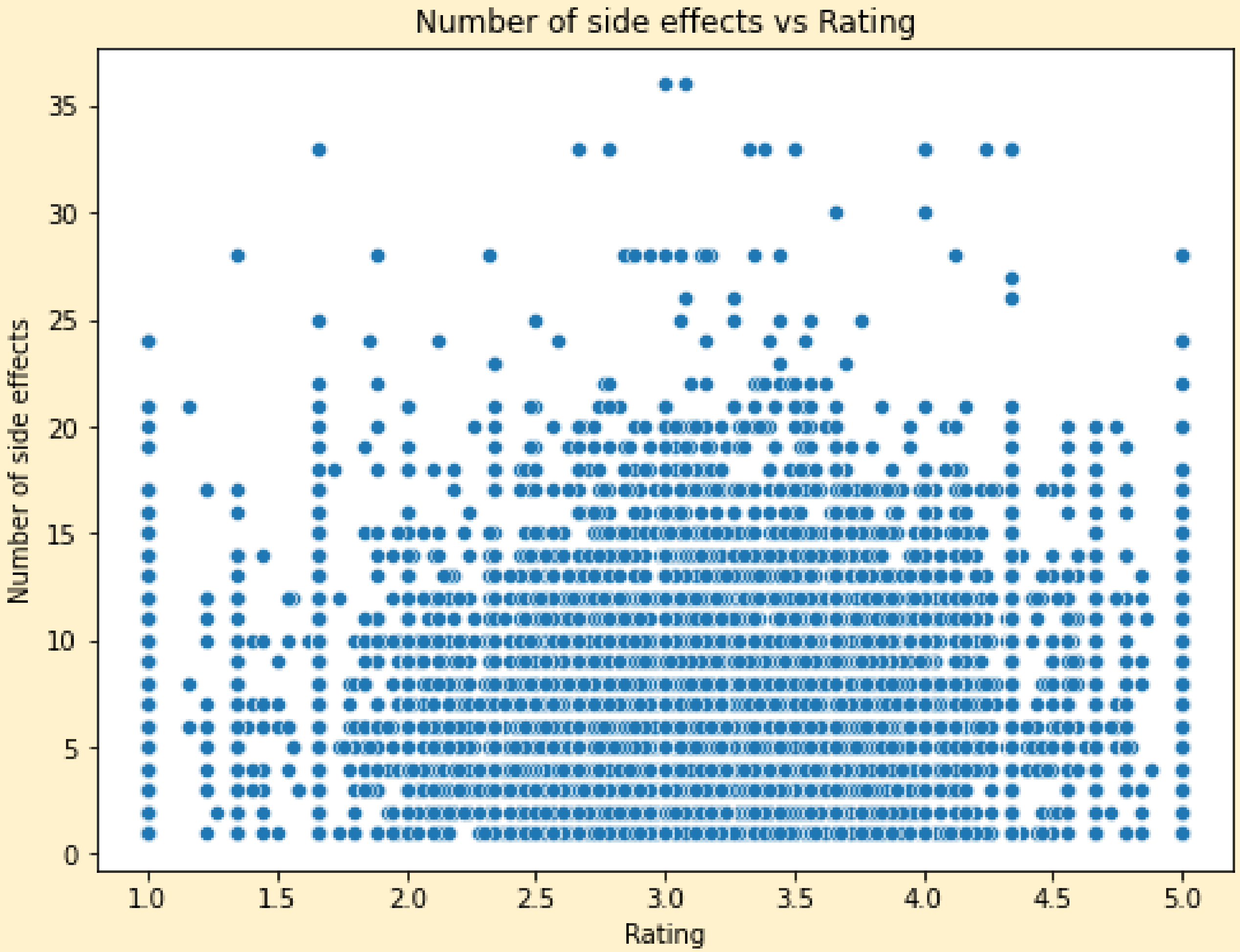


Top 10 Medicines Based On Rating



- Top 10 medicines with rating 5 includes Nigtel 40mg Tablet, ZO Eye/Ear Drops, Pregcert -AQ 25 Injection etc.

Number Of Side Effects Vs Rating For A Medicine

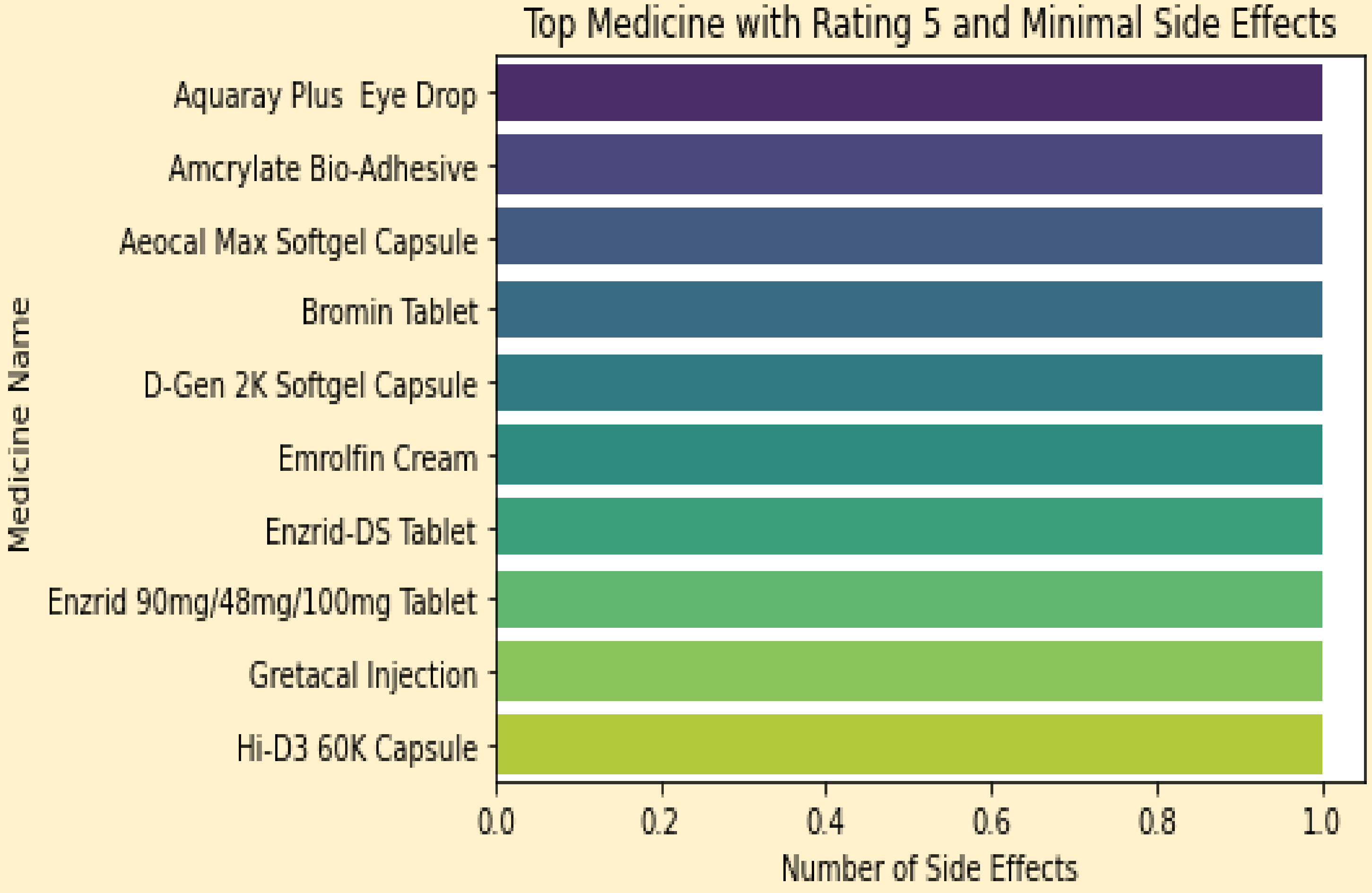


- There is no clear or consistent correlation between the number of side effects and the rating.

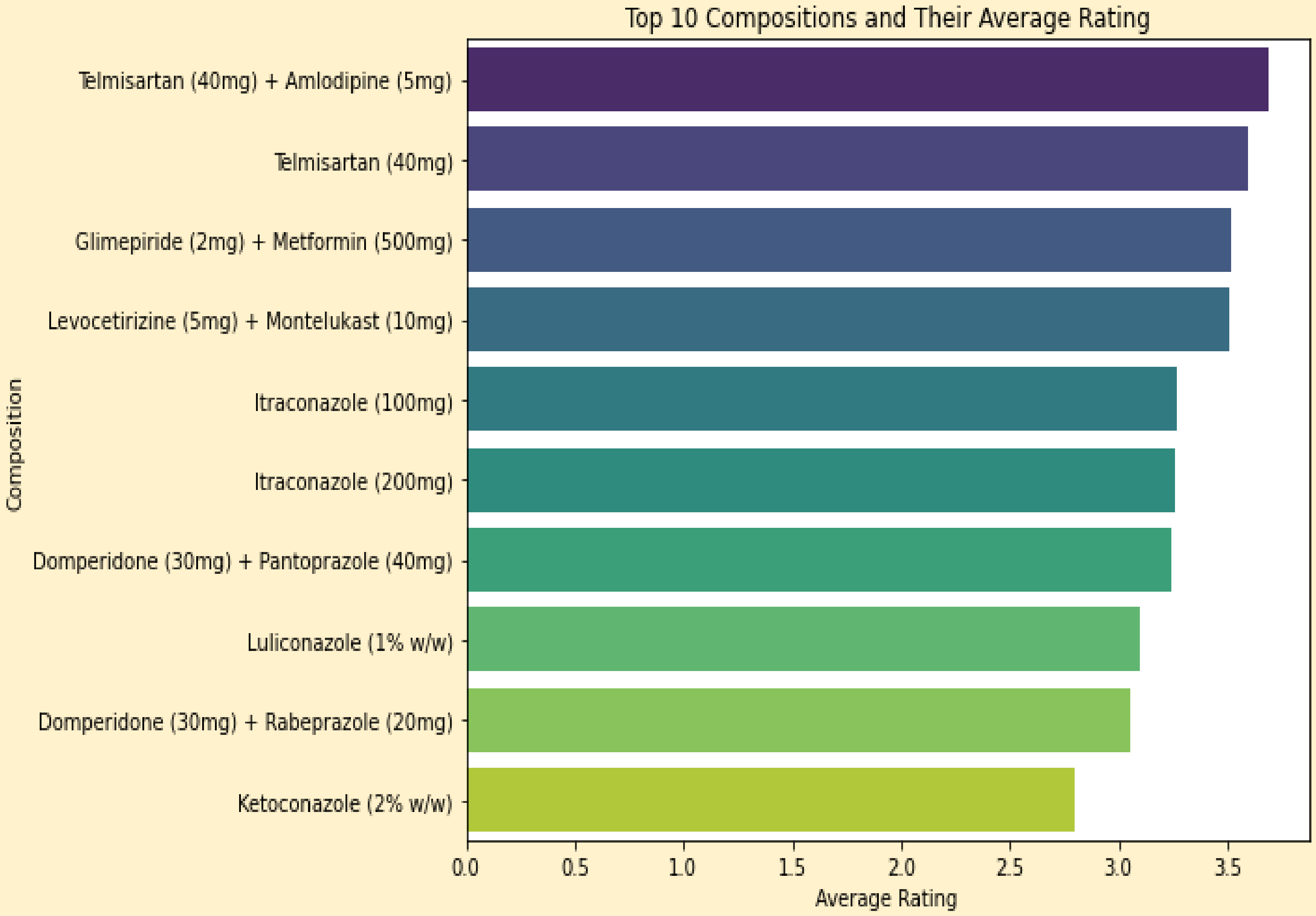
Top 10 Medicine with Rating 5 and Minimal Side Effects

- The top-rated medicines (rating 5) with fewer side effects includes, Aquaray Plus Eye Drop, Amcrylate Bio-Adhesive, Aeocal Max Softgel Capsule etc.

- All these have rating 5 and side effects less than 2.



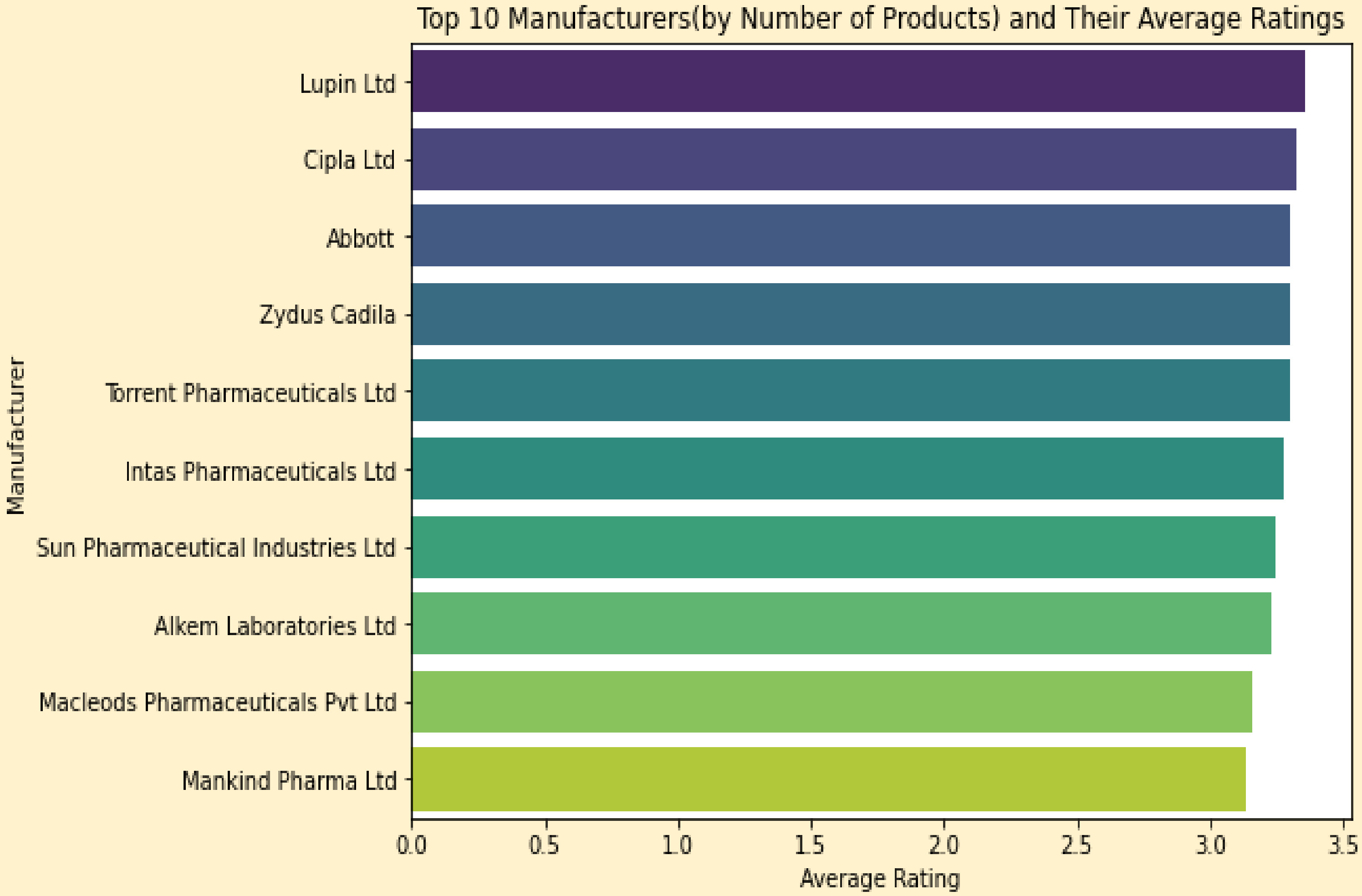
Top 10 Compositions And Their Average Rating



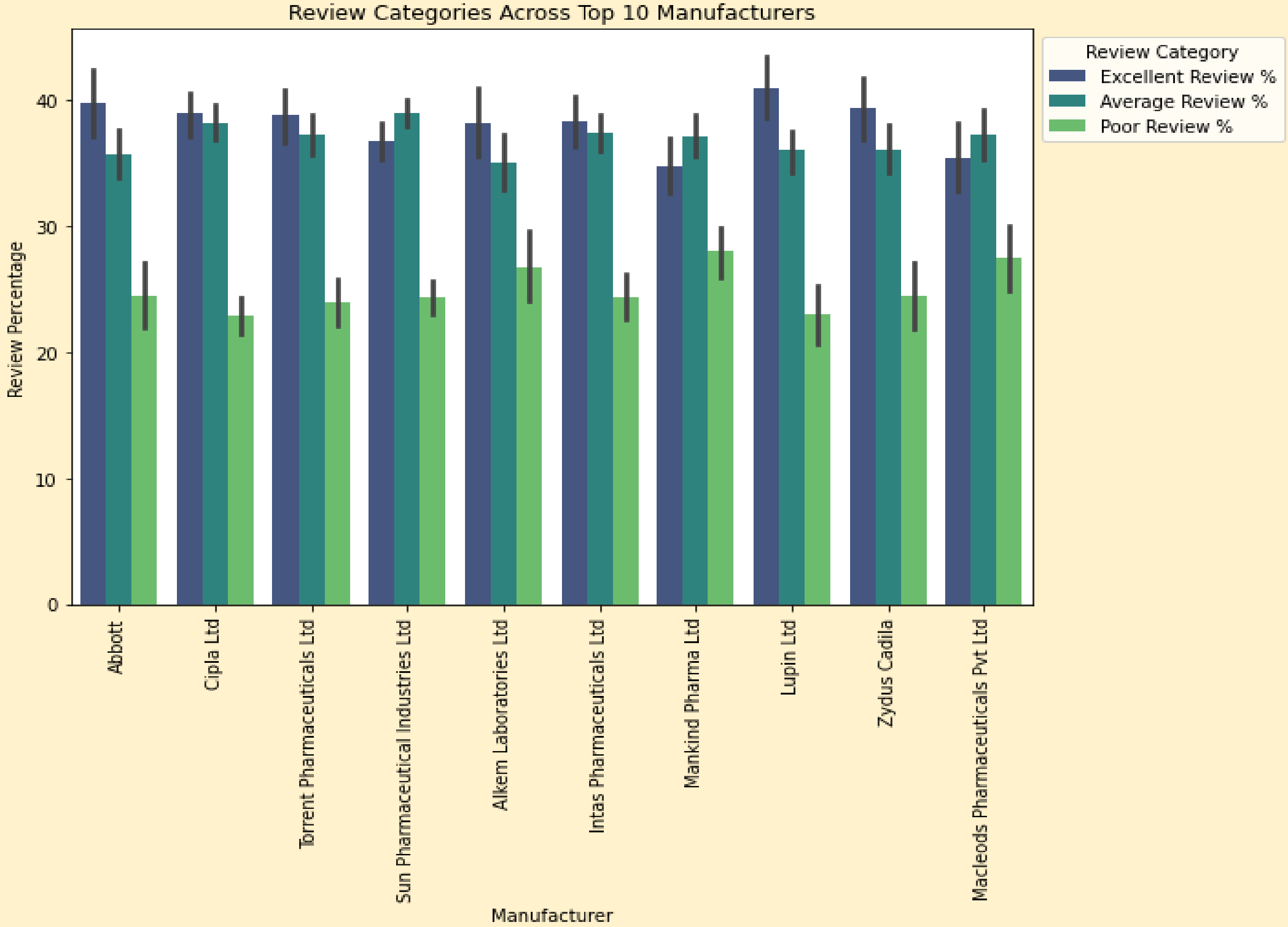
- ❑ **Telmisartan (40mg) + Amlodipine (5mg):** Composition with highest average rating, 3.69
- ❑ **Telmisartan (40mg) :** Composition with second highest average rating ,3.59
- ❑ **Glimepiride (2mg) + Metformin (500mg):** Composition with the third-highest average rating, 3.52

Top 10 Manufactures(Based On Number Of Products) And Their Average Ratings

- **Lupin Ltd**: Manufacturer with the highest average rating, 3.36
- **Cipla Ltd**: Manufacturer with the second-highest average rating, 3.32
- **Abbott**: Manufacturer with the third-highest average rating, 3.30

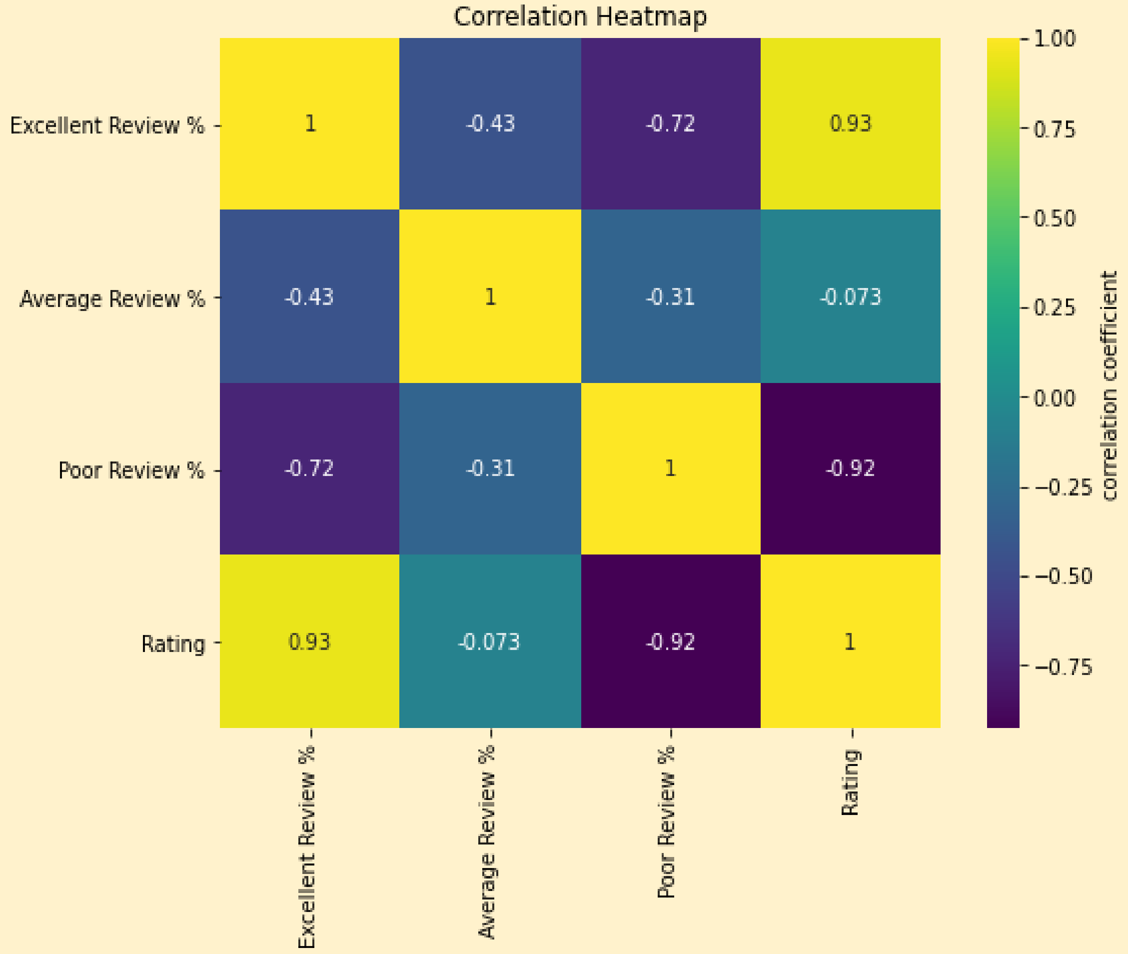


Review Category Of Top 10 Manufactures (Manufactures With Highest Number Of Products)



- All top manufacturers have a higher proportion of Excellent and Average reviews compared to Poor reviews.
- **Lupin Ltd**: Manufacturer with the highest proportion of excellent review, suggesting a strong positive perception of their products among consumers.
- **Abbott:** Also has a notably high percentage of Excellent reviews.

Correlation Between Excellent Review, Average Review, Poor Review, Rating



- ❑ **Excellent Review % and Rating (0.93):** A strong positive correlation exists between Excellent Review and Rating, indicating that a higher percentage of excellent reviews tends to result in a higher rating.
- ❑ **Excellent Review % and Poor Review % (-0.72):** A strong negative correlation exists between Excellent Review and Poor Review, suggesting that an increase in excellent reviews is associated with a decrease in poor reviews.
- ❑ **Poor Review % and Rating (-0.92):** A very strong negative correlation between Poor Review and Rating, indicating that a higher percentage of poor reviews leads to a lower rating.

Feature And Target Set

Feature Set

- Medicine Name
- Manufacturer
- Composition
- Excellent Review %
- Average Review %
- Poor Review %

Target Variable

- Rating

Encoding Techniques

Label Encoder

- Medicine Name
- Manufacturer

TF-IDF

- Composition

- ❑ **Label Encoding**: converts categorical values into numeric labels, simplifying the representation of data for machine learning models.
- ❑ **TF-IDF**: assesses the relevance of a word in a document based on its frequency in the document and its rarity across all documents.

Models And Parameters

KNeighborsRegressor

- n_neighbors=5

DecisionTreeRegressor

- max_depth=3
- min_samples_split=10
- min_samples_leaf=2
- random_state=42

RandomForestRegressor

- n_estimators=100
- max_depth=3
- random_state=42

Model Evaluation

Random Forest

- Mean squared error: 0.029
- R2: 0.964

Decision Tree

- Mean squared error: 0.055
- R2: 0.934

KNN

- Mean squared error: 0.214
- R2: 0.745

Random Forest Regressor: lowest MSE (0.0296) and the highest R2 (0.9647), indicating the best predictive accuracy.

Decision Tree Regressor: shows strong performance with a high R2 (0.9345) and relatively low MSE (0.0550).

Kneighbors Regressor: slightly less effective with an R^2 of 0.745 and a higher MSE of 0.214.

Model Deployment

Deployed the best model (Random Forest Regressor) for real-time predictions using Streamlit

Key Insights

- **Strong Correlations**: The analysis revealed strong correlations between the Excellent Review % and Rating, suggesting that higher percentages of excellent reviews are often associated with higher ratings.

- **Impact of Side Effects**: The relationship between side effects and rating was analyzed, though it was found that the number of side effects does not strongly influence the rating.

- **Top Performers**: Certain compositions and manufacturers consistently received higher ratings, with Telmisartan (40mg) + Amlodipine (5mg) and Lupin Ltd standing out as top performers in terms of average ratings.

- **Review Categories**: Excellent Reviews and Poor Reviews showed a clear inverse relationship, where higher excellent reviews correlate with fewer poor reviews, highlighting the significance of positive feedback in user satisfaction.

- **Prediction**: Medicine Name, Manufacturer, Composition, and Review Categories (Excellent Review %, Average Review %, Poor Review %) were used as key features to predict the user satisfaction rating of medicines.

Thank You!